Mu Yuan

Website: https://yuanmu97.github.io/
Email: muyuan@cuhk.edu.hk
GitHub: github.com/yuanmu97

Education

University of Science and Technology of China — Hefei, China
Ph.D. in Computer Science and Technology — Sep. 2019–Jun. 2024

- Advisor: Prof. Xiang-Yang Li (ACM/IEEE Fellow) and Prof. Lan Zhang
- Dissertation: Heterogeneous Collaborative Model Inference
- USTC Doctoral Dissertation Award

University of Science and Technology of China — Hefei, China
B.S. in Computer Science and Technology (Hua-Xia Talent Class) — Sep. 2015–Jun. 2019

- Thesis: Comprehensive and Efficient Data Labelling via Adaptive Model Scheduling
- USTC Outstanding Undergraduate Thesis Award

Experience

The Chinese University of Hong Kong — Hong Kong, China
AIoT Lab, Postdoctoral Fellow, Prof. Guoliang-Xing (ACM/IEEE Fellow) — Jul. 2024-Current

SenseTime — Beijing, China
Intern Researcher, Deep Learning-Based Video Action Recognition Project — Mar.-Jul. 2019

University of Washington — Seattle, U.S.
Summer Research Program — Jul.-Sep. 2017

Awards and Grants

National Natural Science Foundation of China, Grant No.623B2093, RMB 300,000	2024-2025
ACM SenSys 2025, Best Paper Honorable Mention Award	2025
CCF Doctoral Dissertation Award (10 nationwide)	2024
CCF TCIoT Doctoral Dissertation Award (4 nationwide)	2024
ACM SenSys 2024, Best Demo Runner-up Award	2024
CAS President Award	2024
ByteDance Scholars (13 nationwide)	2023
National Scholarship	2020/2022/2023
SenseTime Scholarship (22 nationwide)	2018
Grand Price (1 out of 1530 teams) of the 4th National University Cloud Computing Contest	2018

Academic Services

- General co-chair of ACM ANAI Workshop 2025 (co-located with ACM MobiCom 2025)
- Reviewer of ACM IMWUT, IEEE INFOCOM, IEEE TMC, IEEE IoTJ, AAAI, NeurIPS

Publications

1. **Mu Yuan**, Lan Zhang, Yihang Cheng, Miao-Hui Song, Guoliang Xing, Xiang-Yang Li. STIP: Three-Party Privacy-Preserving and Lossless Inference for Large Transformers in Production. In The Network and Distributed System Security (NDSS) Symposium. 2026.

2. **Mu Yuan**, Lan Zhang, Liekang Zeng, Siyang Jiang, Bufang Yang, Di Duan, Guoliang Xing. SCX: Stateless KV-Cache Encoding for Cloud-Scale Confidential Transformer Serving. In ACM SIGCOMM Conference. 2025.

3. Liekang Zeng, Yunchao Liu, Shengyuan Ye, **Mu Yuan**, Di Duan, Xu Chen, Guoliang Xing. Grape: Efficient Spatiotemporal Prediction Services with Stale Sensing Streams. In The IEEE Real-Time Systems Symposium (RTSS). 2025.

4. Siyang Jiang, Bufang Yang, Lilin Xu, **Mu Yuan**, Yeerzhati Abudunuer, Kaiwei Liu, Liekang Zeng, Hongkai Chen, Xiaofan Jiang, Zhenyu Yan, Guoliang Xing. LLM-Driven Low-Resolution Vision System for On-Device Human Behavior Understanding. In ACM MobiCom. 2025.

5. Yuting He, Xinyan Wang, **Mu Yuan**, Bufang Yang, Siyang Jiang, Yihua Huang, Doris S. F. Yu, Guoliang Xing, Hongkai Chen. Myo-Trainer: A Vision-based Muscle-Aware Motion Feedback System for In-Home Resistance Training. In ACM MobiCom. 2025.

6. **Mu Yuan**, Lan Zhang, Di Duan, Liekang Zeng, Miao-Hui Song, Zichong Li, Guoliang Xing, Xiang-Yang Li. Mitigating Tail Latency for on-Device Inference with Load-Balanced Heterogeneous Models. In IEEE Transactions on Mobile Computing. 2025.

7. Yihang Cheng, Lan Zhang, Junyang Wang, **Mu Yuan**, Yunhao Yao. RemoteRAG: A Privacy-Preserving LLM Cloud RAG Service. In Findings of the Association for Computational Linguistics (ACL). 2025.

8. Puhan Luo, Jiahui Hou, Haisheng Tan, **Mu Yuan**, Guangyu Wu, Kaiwen Guo, Zhiqiang Wang, XiangYang Li. ChannelZip: SLO-aware channel compression for task-adaptive model serving on IoT devices. In ACM Transactions on Sensor Networks. 2025.

9. Di Duan, Shengzhe Lyu, **Mu Yuan**, Hongfei Xue, Tianxing Li, Weitao Xu, Kaishun Wu, Guoliang Xing. Argus: Multi-view egocentric human mesh reconstruction based on stripped-down wearable mmwave add-on. In Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems. 2025.

10. Junyang Zhang, **Mu Yuan**, Ruiguang Zhong, Puhan Luo, Huiyou Zhan, Ningkang Zhang, Chengchen Hu, Xiang-Yang Li. A-VL: Adaptive Attention for Large Vision-Language Models. In Proceedings of the AAAI Conference on Artificial Intelligence. 2025.

11. Yunhao Yao, Jiahui Hou, **Mu Yuan**, Haiyue Zhang, Zhengyuan Xu, Xiang-Yang Li. TrafficDiary: User Attribute Inference Based on Smart Home Traffic Traces. In ACM Transactions on Internet Technology. 2025.

12. **Mu Yuan**, Lan Zhang, Yunhao Yao, Junyang Zhang, Puhan Luo, Xiang-Yang Li. Resource-Efficient Model Inference for AIoT: A Survey. In the Chinese Journal of Computers. 2024.

13. Yunhao Yao, Jiahui Hou, Guangyu Wu, Yihang Cheng, **Mu Yuan**, Puhan Luo, Zhiqiang Wang, Xiang-Yang Li. SecoInfer: Secure DNN End-Edge Collaborative Inference Framework Optimizing Privacy and Latency. In ACM Transactions on Sensor Networks. 2024.

14. Puhan Luo, Jiahui Hou, **Mu Yuan**, Guangyu Wu, Yunhao Yao, Xiang-Yang Li. F2Zip: Finetuning-free model compression for scenario-adaptive embedded vision. In Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems. 2024.

15. Yuting He, Xinyan Wang, **Mu Yuan**, Di Duan, Doris SF Yu, Guoliang Xing, Hongkai Chen. Demo: Myotrainer: Muscle-Aware Motion Analysis and Feedback System for In-Home Resistance Training. In Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems. 2024.

16. Ningkang Zhang, Guangyu Wu, Chao Gu, **Mu Yuan**, Xiang-Yang Li. FusionFlow: Neural Fusion and Compression for Communication-Efficient Edge-Cloud Collaborative Computing. In International Conference on Wireless Artificial Intelligent Computing Systems and Applications. 2024.

17. Junyang Wang, Lan Zhang, Junhao Wang, **Mu Yuan**, Yihang Cheng, Qian Xu, Bo Yu. GraphProxy: Communication-efficient federated graph learning with adaptive proxy. In IEEE INFOCOM. 2024.

18. **Mu Yuan**, Lan Zhang, Xuanke You, and Xiang-Yang Li. PacketGame: Multi-Stream Packet Gating for Concurrent Video Inference at Scale. In ACM SIGCOMM Conference. 2023.

19. **Mu Yuan**, Lan Zhang, Fengxiang He, Xueting Tong, Zhenyuan Xu, and Xiang-Yang Li. InFi: End-to-End Learning to Filter Input for Resource-Efficiency in Mobile-Centric Inference. In IEEE Transactions on Mobile Computing (TMC). 2023.

20. **Mu Yuan**, Lan Zhang, Zimu Zheng, Yi-Nan Zhang, and Xiang-Yang Li. MLink: Linking Black-Box Models from Multiple Domains for Collaborative Inference. In IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2023.

21. Miao-Hui Song, Lan Zhang, **Mu Yuan**, Zichong Li, Qi Song, Yijun Liu, Guidong Zheng. Cotel: Ontology-neural co-enhanced text labeling. In Proceedings of the ACM Web Conference. 2023.

22. Zichong Li, Lan Zhang, **Mu Yuan**, Miao-Hui Song, and Qi Song. Efficient Deep Ensemble Inference via Query Difficulty-dependent Task Scheduling. In IEEE ICDE Conference. 2023.

23. Lan Zhang, Daren Zheng, **Mu Yuan**, Feng Han, Zhengtao Wu, Mengjing Liu, and Xiang-Yang Li. MultiSense: Cross-labelling and Learning Human Activities Using Multimodal Sensing Data. In ACM Transactions on Sensor Networks (TOSN). 2023.

24. **Mu Yuan**, Lan Zhang, Fengxiang He, Xueting Tong, and Xiang-Yang Li. InFi: End-to-End Learnable Input Filter for Resource-Efficient Mobile-Centric Inference. In **ACM MobiCom** Conference. 2022.

25. **Mu Yuan**, Lan Zhang, Xiang-Yang Li, Lin-Zhuo Yang, and Hui Xiong. Adaptive Model Scheduling for Resource-efficient Data Labeling. In ACM Transactions on Knowledge Discovery from Data (TKDD). 2022.

26. **Mu Yuan**, Lan Zhang, and Xiang-Yang Li. MLink: Linking Black-Box Models for Collaborative Multi-Model Inference. In AAAI Conference. 2022. (Oral Presentation 4.5%)

27. Xuanke You, Lan Zhang, Haikuo Yu, **Mu Yuan**, and Xiang-Yang Li. KATN: Key activity detection via inexact supervised learning. In ACM Ubicomp Conference. 2021.

28. Lan Zhang, **Mu Yuan**, Daren Zheng, Xiang-Yang Li. M&M: Recognizing Multiple Co-evolving Activities from Multi-Source Videos. In 17th International Conference on Distributed Computing in Sensor Systems (DCOSS). 2021.

29. **Mu Yuan**, Lan Zhang, Zhengtao Wu, and Daren Zheng. High-quality Activity-Level Video Advertising. In IEEE/ACM IWQoS Conference. 2020.

30. **Mu Yuan**, Lan Zhang, Xiang-Yang Li, and Hui Xiong. Comprehensive and Efficient Data Labeling via Adaptive Model Scheduling. In IEEE ICDE Conference. 2020.